Nested Hierarchical Dirichlet Processes

John Paisley[1], Chong Wang[3], David M. Blei[4] and Michael I. Jordan[1,2]

[1]Department of EECS, [2]Department of Statistics, UC Berkeley, Berkeley, CA

[3]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA

[4]Department of Computer Science, Princeton University, Princeton, NJ

Abstract

We develop a nested hierarchical Dirichlet process (nHDP) for hierarchical topic modeling. The nHDP is a generalization of the nested Chinese restaurant process (nCRP) that allows each word to follow its own path to a topic node according to a document-specific distribution on a shared tree. This alleviates the rigid, single-path formulation of the nCRP, allowing a document to more easily express thematic borrowings as a random effect. We derive a stochastic variational inference algorithm for the model, in addition to a greedy subtree selection method for each document, which allows for efficient inference using massive collections of text documents. We demonstrate our algorithm on 1.8 million documents from *The New York Times* and 3.3 million documents from *Wikipedia*.

Index Terms

Bayesian nonparametrics, Dirichlet process, topic modeling, stochastic inference

I. Introduction

Organizing things hierarchically is a natural process of human activity. Walking into a large department store, one might first find the men's section, followed by men's casual, and then see the t-shirts hanging along the wall. Or one may be in the mood for Italian food, decide whether to spring for the better, more authentic version or go to one of the cheaper chain options, and then end up at the Olive Garden. Similarly with data analysis, a hierarchical tree-structured representation of the data can provide an illuminating means for understanding and reasoning about the information it contains.

The nested Chinese restaurant process (nCRP) [1] is a model that performs this task for the problem of topic modeling. *Hierarchical topic models* place a structured prior on the topics underlying a corpus of documents, with the aim of bringing more order to an unstructured set of thematic concepts [1][2][3].

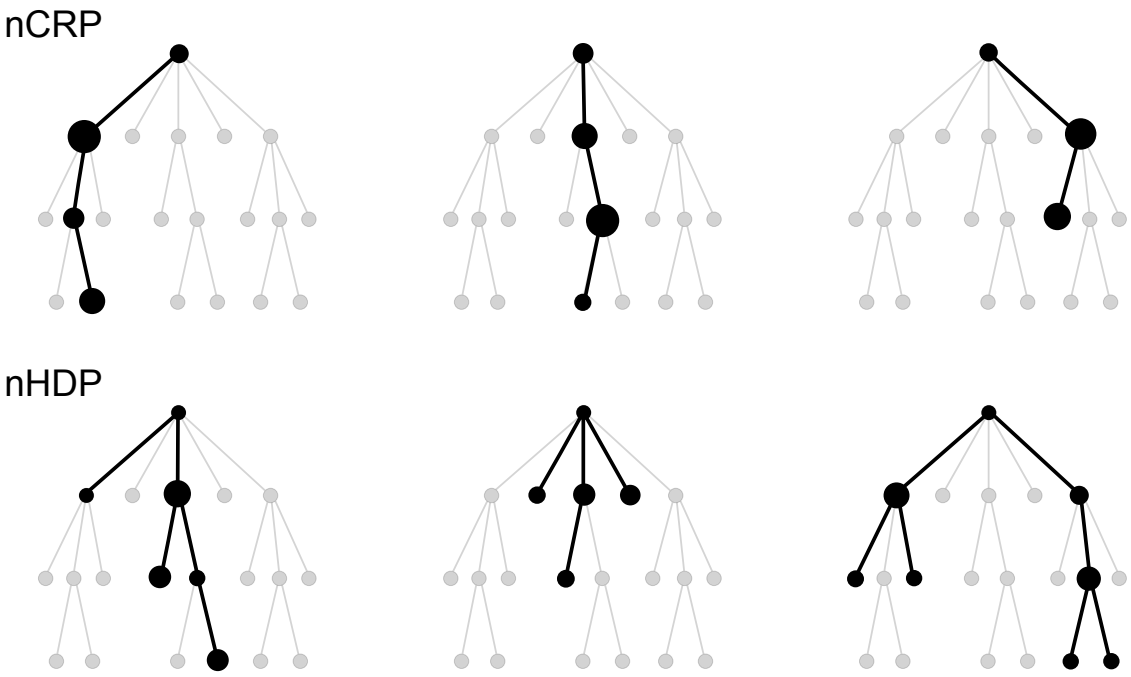

Fig. 1. An example of path structures for the nested Chinese restaurant process (nCRP) and the nested hierarchical Dirichlet process (nHDP) for hierarchical topic modeling. With the nCRP, the topics for a document are restricted to lying along a single path to a root node. With the nHDP, each document has access to the entire tree, but a document-specific distribution on paths will place high probability on a particular subtree. The goal of the nHDP is to learn a thematically consistent tree as achieved by the nCRP, while allowing for the cross-thematic borrowings that naturally occur within a document.

They do this by learning a tree structure for the underlying topics, with the inferential goal being that topics closer to the root are more general, and gradually become more specific in thematic content when following a path down the tree.

The nCRP is a Bayesian nonparametric prior for hierarchical topic models, but is limited in the hierarchies it can model. We illustrate this limitation in Figure 1. The nCRP models the topics that go into constructing a document as lying along one path of the tree. From a practical standpoint this is a disadvantage, since inference in trees over three levels is computationally hard [2][1], and hence in practice each document is limited to only three underlying topics. Moreover, this is also a significant disadvantage from a modeling standpoint.

As a simple example, consider a document on ESPN.com about an injured player, compared with an article in a sports medicine journal. Both documents will contain words about medicine and words about sports. Should the nCRP select a path transitioning from sports to medicine, or vice versa? Depending on the article, both options are reasonable, and during the learning process the model will either acquire

both paths, hence partitioning sports and medicine words among multiple topics, or choose one over the other, which will require all documents containing the topic from the lower level to least have the higher level topic activated. In one case the model is not using the full statistical power within the corpus to model each topic and in the other the model is learning an unreasonable tree. Returning to the practical aspect, for trees truncated to a small number of levels, there simply is not enough room to learn all of these combinations.

Though the nCRP is a Bayesian nonparametric prior, it performs nonparametric clustering of *document-specific* paths, which fixes the number of available topics to a small number for trees of a few levels. Our goal is to develop a related Bayesian nonparametric prior that performs *word-specific* path clustering. We illustrate this objective in Figure 1. In this case, each word has access to the entire tree, but with document-specific distributions on paths within the tree. To this end, we make use of the hierarchical Dirichlet process [4], developing a novel prior that we refer to as the *nested hierarchical Dirichlet process* (nHDP). The HDP can be viewed as a nonparametric elaboration of the classical topic model, the latent Dirichlet allocation (LDA) model [5], providing a mechanism whereby a top-level Dirichlet process provides a base distribution for a collection of second-level Dirichlet processes, one for each document. With the nHDP, a top-level nCRP becomes a base distribution for a collection of second-level nCRPs, one for each document. The nested HDP provides the opportunity for cross-thematic borrowing that is not possible with the nCRP.

Hierarchical topic models have thus far been applied to corpora of small size. A significant issue, not just with topic models but with Bayesian models in general, is scaling up inference to massive data sets [6]. Recent developments in stochastic variational inference methods have done this for LDA and the HDP topic model [7][8][9]. We continue this development for hierarchical topic modeling with the nested HDP. Using stochastic VB, in which we maximize the variational objective using stochastic optimization, we demonstrate the ability to efficiently handle very large corpora. This is a major benefit to complex models such as tree-structured topic models, which require significant amounts of data to support their exponential growth in size.

We organize the paper as follows: In Section II we review the Bayesian nonparametric priors that we incorporate in our model—the Dirichlet process, nested Chinese restaurant process and hierarchical Dirichlet process. In Section III we present our proposed nested HDP model for hierarchical topic modeling. In Section IV we review stochastic variational inference and present an inference algorithm for nHDPs that scales well to massive data sets. We present empirical results in Section V. We first compare the nHDP with the nCRP on three relatively small data sets. We then evaluate our stochastic algorithm on

1.8 million documents from *The New York Times* and 3.3 million documents from *Wikipedia*, comparing performance with stochastic LDA and stochastic HDP.

II. Background: Bayesian nonparametric priors for topic models

The nested hierarchical Dirichlet process (nHDP) builds on a collection of existing Bayesian nonparametric priors. In this section, we provide a review of these priors: the Dirichlet process, nested Chinese restaurant process and hierarchical Dirichlet process. We also review constructive representations for these processes that we will use for posterior inference of the nHDP topic model.

A. Dirichlet processes

The Dirichlet process (DP) [10] is the foundation for a large collection of Bayesian nonparametric models that rely on mixtures to statistically represent data. Mixture models work by partitioning a data set according to statistical traits shared by members of the same cell. Dirichlet process priors are effective in the learning of the number of these traits, in addition to the parameters of the mixture. The basic form of a Dirichlet process mixture model is

$$W_n|\varphi_n \sim F_W(\varphi_n), \quad \varphi_n|G \overset{iid}{\sim} G, \quad G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}. \tag{1}$$

With this representation, data $W_1, \ldots, W_N$ are distributed according to a family of distributions F_W with respective parameters $\varphi_1, \ldots, \varphi_N$. These parameters are drawn from the distribution G, which is discrete and potentially infinite, as the DP allows it to be. This discreteness induces a partition of the data W according to the sharing of the atoms $\{\theta_i\}$ among the parameter selections $\{\varphi_n\}$.

The Dirichlet process is a stochastic process on random elements G. To briefly review, let $(\Theta, \mathcal{B})$ be a measurable space, G_0 a probability measure on it and $\alpha > 0$. Ferguson proved the existence of a stochastic process G where, for all partitions $\{B_1, \ldots, B_k\}$ of Θ,

$$(G(B_1), \ldots, G(B_k)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_k)),$$

abbreviated as $G \sim \text{DP}(\alpha G_0)$. It has been shown that G is discrete (with probability one) even when G_0 is non-atomic [11][12], though the probability that the random variable $G(B_k)$ is less than ϵ increases to 1 as B_k decreases to a point for every $\epsilon > 0$. Thus the DP prior is a good candidate for G in (1) since it generates discrete distributions on infinitely large parameter spaces. For most applications G_0 is continuous, and so representations of G at the granularity of the atoms are necessary for inference; we next review two approaches to working with this infinite-dimensional distribution.

1) Chinese restaurant process: The Chinese restaurant process (CRP) avoids directly working with G by integrating it out [11][13]. In doing so, the values of $\varphi_1, \ldots, \varphi_N$ become dependent, with the value of φ_{n+1} given $\varphi_1, \ldots, \varphi_n$ distributed as

$$\varphi_{n+1}|\varphi_1, \ldots, \varphi_n \sim \sum_{i=1}^{n} \frac{1}{\alpha+n}\delta_{\varphi_i} + \frac{\alpha}{\alpha+n}G_0. \tag{2}$$

That is, φ_{n+1} takes the value of one of the previously observed φ_i with probability $\frac{n}{\alpha+n}$, and a value drawn from G_0 with probability $\frac{\alpha}{\alpha+n}$, which will be unique when G_0 is continuous. This displays the clustering property of the CRP and also gives insight into the impact of α, since it is evident that the number of unique φ_i grows like $\alpha \ln(\alpha + n)$. In the limit $n \to \infty$, the distribution in (2) converges to a random measure distributed according to a Dirichlet process [11]. The CRP is so-called because of an analogy to a Chinese restaurant, where a customer (datum) sits at a table (selects a parameter) with probability proportional to the number of previous customers at that table, or selects a new table with probability proportional to α.

2) Stick-breaking construction: Where the Chinese restaurant process works with $G \sim \text{DP}(\alpha G_0)$ implicitly through φ, a stick-breaking construction allows one to directly construct G before drawing any φ_n. Sethuraman [12] showed that if G is constructed as follows:

$$G = \sum_{i=1}^{\infty} V_i \prod_{j=1}^{i-1}(1 - V_j)\delta_{\theta_i}, \quad V_i \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad \theta_i \overset{iid}{\sim} G_0, \tag{3}$$

then $G \sim \text{DP}(\alpha G_0)$. The variable V_i can be interpreted as the proportion broken from the remainder of a unit length stick, $\prod_{j<i}(1 - V_j)$. As the index i increases, more random variables in $[0, 1]$ are multiplied, and thus the weights exponentially decrease to zero; the expectation $\mathbb{E}[V_i \prod_{j<i}(1 - V_j)] = \frac{\alpha^{i-1}}{(1+\alpha)^i}$ gives a sense of the impact of α on these weights. This explicit construction of G maintains the independence among $\varphi_1, \ldots, \varphi_N$ as written in Equation (1), which is a significant advantage of this representation for mean-field variational inference that is not present in the CRP.

B. Nested Chinese restaurant processes

Nested Chinese restaurant processes (nCRP) are a tree-structured extension of the CRP that are useful for hierarchical topic modeling [1]. They extend the CRP analogy to a nesting of restaurants in the following way: After selecting a table (parameter) according to a CRP, the customer departs for another restaurant only indicated by that table. Upon arrival, the customer again acts according to the CRP for the new restaurant, and again departs for a restaurant only accessible through the table selected. This

occurs for a potentially infinite sequence of restaurants, which generates a sequence of parameters for the customer according to the selected tables.

A natural interpretation of the nCRP is as a tree where each parent has an infinite number of children. Starting from the root node, a path is traversed down the tree. Given the current node, a child node is selected with probability proportional to the previous number of times it was selected among its siblings, or a new child is selected with probability proportional to α. As with the CRP, the nCRP also has a constructive representation useful for variational inference which we now discuss.

1) Constructing the nCRP: The nesting of Dirichlet processes that leads to the nCRP gives rise to a stick-breaking construction [2].[1] We develop the notation for this construction here and use it later in our construction of the nested HDP. Let $\boldsymbol{i}_l = (i_1, \dots, i_l)$ be a path to a node at level l of the tree.[2] According to the stick-breaking version of the nCRP, the children of node $\boldsymbol{i}_l$ are countably infinite, with the probability of transitioning to child j equal to the jth break of a stick-breaking construction. Each child corresponds to a parameter drawn independently from G_0. Letting the index of the parameter identify the index of the child, this results in the following DP for the children of node $\boldsymbol{i}_l$,

$$G_{\boldsymbol{i}_l} = \sum_{j=1}^{\infty} V_{\boldsymbol{i}_l,j} \prod_{m=1}^{j-1} (1 - V_{\boldsymbol{i}_l,m}) \delta_{\theta_{(\boldsymbol{i}_l,j)}}, \quad V_{\boldsymbol{i}_l,j} \overset{iid}{\sim} \text{Beta}(1, \alpha), \quad \theta_{(\boldsymbol{i}_l,j)} \overset{iid}{\sim} G_0. \tag{4}$$

If the next node is child j, then the nCRP transitions to DP $G_{\boldsymbol{i}_{l+1}}$, where $\boldsymbol{i}_{l+1}$ has index j appended to $\boldsymbol{i}_l$, that is $\boldsymbol{i}_{l+1} = (\boldsymbol{i}_l, j)$. A sequence of parameters $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \dots)$ generated from a path down this tree follows a Markov chain, where the parameter φ_l correspond to an atom $\theta_{\boldsymbol{i}_l}$ at level l and the stick-breaking weights correspond to the transition probabilities. Hierarchical topic models use these sequences of parameters as topics for generating documents.

2) Nested CRP topic models: Hierarchical topic models based on the nested CRP use a globally shared tree to generate a corpus of documents. Starting with the construction of nested Dirichlet processes as described above, each document selects a path down the tree according to a Markov process, which produces a sequence of topics $\boldsymbol{\varphi}_d = (\varphi_{d,1}, \varphi_{d,2}, \dots)$ used to generate the document. As with other topic models, each word in a document is represented by an index $W_{d,n} \in \{1, \dots, \mathcal{V}\}$ and the atoms $\theta_{\boldsymbol{i}_l}$ appearing in $\boldsymbol{\varphi}_d$ are $\mathcal{V}$-dimensional probability vectors with prior G_0 a Dirichlet distribution.

[1]The "nested Dirichlet process" that we present here was first described (using random measures rather than the stick-breaking construction) by [14], who developed it for a two-level tree.

[2]That is, from the root node first select the child with index i_1; from node $\boldsymbol{i}_1 = (i_1)$, select the child with index i_2; from node $\boldsymbol{i}_2 = (i_1, i_2)$ select the child with index i_3, and so on to level l. We ignore the root $\boldsymbol{i}_0$, which is shared by all paths.

For each document d, a new stick-breaking process provides a distribution on the topics in $\boldsymbol{\varphi}_d$,

$$G^{(d)} = \sum_{j=1}^{\infty} U_{d,j} \prod_{m=1}^{j-1} (1 - U_{d,m}) \delta_{\varphi_{d,j}}, \quad U_{d,j} \overset{iid}{\sim} \text{Beta}(\gamma_1, \gamma_2). \tag{5}$$

Following the standard method, words for document d are generated by first drawing a parameter i.i.d. from $G^{(d)}$, and then drawing the word index from the discrete distribution with the selected parameter.

3) Issues with the nCRP: As discussed in the introduction, a significant drawback of the nCRP for topic modeling is that each document follows one path down the tree. Therefore, all thematic content of a document must be contained within that single sequence of topics. Since the nCRP is meant to characterize the thematic content of a corpus in increasing levels of specificity, this creates a combinatorial problem, where similar topics will appear in many parts of the tree to account for the possibility that they appear as a random effect in a document. In practice, nCRP trees are typically truncated at three levels [2][1], since learning deeper levels becomes difficult due to the exponential increase in nodes.[3] In this situation each document has three topics for modeling its entire thematic content, and so a blending of multiple topics is likely to occur during inference.

The nCRP is a BNP prior, but it performs nonparametric clustering of the paths selected at the document level, rather than at the word level. Though the same tree is shared by a corpus, each document can differentiate itself by the path it choses. The key issue with the nCRP is the restrictiveness of this single path allowed to a document. If instead each word were allowed to follow its own path according to an nCRP, this characteristic would be lost and only a tree level distribution similar to Equation (5) could distinguish one document from another and thematic coherence would be missing. Our goal is to develop a hierarchical topic model that does not prohibit a document from using topics in different parts of the tree. Our solution to this problem is to employ the hierarchical Dirichlet process (HDP) [4].

C. Hierarchical Dirichlet processes

The HDP is a multi-level version of the Dirichlet process. It makes use of the idea that the base distribution on the infinite space Θ can be discrete, and indeed a discrete distribution allows for multiple draws from the DP prior to place probability mass on the same subset of atoms. Hence different groups of data can share the same atoms, but place different probability distributions on them. A discrete base is needed, but the atoms are unknown in advance. The HDP achieves this by drawing the base from a

[3]This includes a root node topic, which is shared by all documents and is intended to collect stop words.

DP prior. This leads to the hierarchical process

$$G_d|G \sim \text{DP}(\beta G), \quad G \sim \text{DP}(\alpha G_0), \tag{6}$$

for groups $d = 1, \ldots, D$. This prior has been used to great effect in topic modeling as a nonparametric extension of LDA [5] and related LDA-based models [15][16][17].

As with the DP, concrete representations of the HDP are necessary for inference. The representation we use relies on two levels of Sethuraman's stick breaking construction. For this construction, after sampling G as in Equation (3), we sample G_d in the same way,

$$G_d = \sum_{i=1}^{\infty} V_i^d \prod_{j=1}^{i-1}(1 - V_j^d)\delta_{\phi_i}, \quad V_i^d \overset{iid}{\sim} \text{Beta}(1, \beta), \quad \phi_i \overset{iid}{\sim} G. \tag{7}$$

This form is identical to Equation (3), with the key difference that G is discrete, and so atoms ϕ_i will repeat. An advantage of this representation is that all random variables are i.i.d., with significant benefits to variational inference strategies.

III. Nested hierarchical Dirichlet processes for topic modeling

In building on the nCRP framework, our goal is to allow for each document to have access to the entire tree, while still learning document-specific distributions on topics that are thematically coherent. Ideally, each document will still exhibit a dominant path corresponding to its main themes, but with offshoots allowing for random effects. Our two major changes to the nCRP formulation toward this end are that (i) each word follows its own path to a topic, and (ii) each document has its own distribution on paths in a shared tree. The BNP tools discussed above make this a straightforward task.

We split the process of generating a document's distribution on topics into two parts: generating a document's distribution on paths down the tree, and generating a word's distribution on terminating at a particular node within those paths.

A. Constructing the tree for a document

With the nHDP, all documents share a global nCRP constructed with a stick-breaking construction as in Section II-B1. Denote this tree by $\mathcal{T}$. As discussed, $\mathcal{T}$ is simply an infinite collection of Dirichlet processes with a continuous base distribution G_0 and a transition rule between DPs. According to this rule, from a root Dirichlet process $G_{\boldsymbol{i}_0}$, a path is followed by drawing $\varphi_{l+1} \sim G_{\boldsymbol{i}_l}$ for $l = 0, 1, 2, \ldots$, where $\boldsymbol{i}_0$ is a constant root index that we ignore, and $\boldsymbol{i}_l = (i_1, \ldots, i_l)$ indexes the current DP associated with $\varphi_l = \theta_{\boldsymbol{i}_l}$. With the nested HDP, we do not perform this path selection on the top-level $\mathcal{T}$, but instead use each Dirichlet process in $\mathcal{T}$ as a base for a second level DP drawn independently for each document.

That is, for document d we construct a tree $\mathcal{T}_d$, where for each $G_{\boldsymbol{i}_l} \in \mathcal{T}$, we draw a corresponding $G_{\boldsymbol{i}_l}^{(d)} \in \mathcal{T}_d$ independently in d according to a second-level Dirichlet process

$$G_{\boldsymbol{i}_l}^{(d)} \sim \text{DP}(\beta G_{\boldsymbol{i}_l}). \tag{8}$$

As discussed in Section II-C, $G_{\boldsymbol{i}_l}^{(d)}$ will have the same atoms as $G_{\boldsymbol{i}_l}$, but with different probability weights on them. Therefore, the tree $\mathcal{T}_d$ will have the same nodes as $\mathcal{T}$, but the probability of a path in $\mathcal{T}_d$ will vary with d, giving each document its own distribution on a shared tree.

We represent this second-level DP with a stick-breaking construction as in Section II-C,

$$G_{\boldsymbol{i}_l}^{(d)} = \sum_{j=1}^{\infty} V_{\boldsymbol{i}_l,j}^{(d)} \prod_{m=1}^{j-1} (1 - V_{\boldsymbol{i}_l,m}^{(d)}) \delta_{\phi_{\boldsymbol{i}_l,j}^{(d)}}, \quad V_{\boldsymbol{i}_l,j}^{(d)} \overset{iid}{\sim} \text{Beta}(1,\beta), \quad \phi_{\boldsymbol{i}_l,j}^{(d)} \overset{iid}{\sim} G_{\boldsymbol{i}_l}. \tag{9}$$

This representation retains full independence among random variables, and will lead to a simple stochastic variational inference algorithm. We note that the atoms from the top-level DP are randomly permuted and copied with this construction; $\phi_{\boldsymbol{i}_l,j}^{(d)}$ does not correspond to the node with parameter $\theta_{(\boldsymbol{i}_l,j)}$. To find the probability mass $G_{\boldsymbol{i}_l}^{(d)}$ places on $\theta_{(\boldsymbol{i}_l,j)}$, one can calculate

$$G_{\boldsymbol{i}_l}^{(d)}(\{\theta_{(\boldsymbol{i}_l,j)}\}) = \textstyle\sum_m G_{\boldsymbol{i}_l}^{(d)}(\{\phi_{\boldsymbol{i}_l,m}^{(d)}\}) \mathbb{I}(\phi_{\boldsymbol{i}_l,m}^{(d)} = \theta_{(\boldsymbol{i}_l,j)}).$$

Using a nesting of HDPs to construct $\mathcal{T}_d$, each document has a tree with transition probabilities defined over the same subset of nodes since $\mathcal{T}$ is discrete, but with values for these probabilities that are document specific. To see how this permits each word to follow its own path while still retaining thematic coherence within a document, consider each $G_{\boldsymbol{i}_l}^{(d)}$ when β is small. In this case, most of the probability will be placed on one atom selected from $G_{\boldsymbol{i}_l}$ since the first proportion $V_{\boldsymbol{i}_l,1}^{(d)}$ will be large with high probability. This will leave little probability remaining for other atoms, a feature of the prior on all second-level DPs in $\mathcal{T}_d$. Starting from the root node of $\mathcal{T}_d$, each word will be highly "encouraged" to select one particular atom at any given node, with some probability of diverging into a random effect topic. In the limit $\beta \to 0$, each $G_{\boldsymbol{i}_l}^{(d)}$ will be a delta function on a $\phi_{\boldsymbol{i}_l,j}^{(d)} \sim G_{\boldsymbol{i}_l}$, and the same path will be selected by each word with probability one, thus recovering the nCRP.

B. Generating a document

With the tree $\mathcal{T}_d$ for document d we have a method for selecting word-specific paths that are thematically coherent. We next discuss generating a document with this tree. As discussed in Section II-B2, with the nCRP the atoms selected for a document by its path through $\mathcal{T}$ have a unique stick-breaking distribution determining which level any particular word comes from. We generalize this idea to the tree $\mathcal{T}_d$ with an overlapping stick-breaking construction as follows.

Algorithm 1 Generating Documents with the Nested Hierarchical Dirichlet Process

Step 1. Generate a global tree $\mathcal{T}$ by constructing an nCRP as in Section II-B1.

Step 2. Generate document tree $\mathcal{T}_d$ and switching probabilities $\boldsymbol{U}^{(d)}$. For document d,

a) For each DP in $\mathcal{T}$, draw a second-level DP with this base distribution (Equation 8).

b) For each node in $\mathcal{T}_d$ (equivalently $\mathcal{T}$), draw a beta random variable (Equation 10).

Step 3. Generate the documents. For word n in document d,

a) Sample atom $\varphi_{n,d} = \theta_{\boldsymbol{i}_l}$ with probability given in Equation (11).

b) Sample $W_{n,d}$ from the discrete distribution with parameter $\varphi_{d,n}$.

For each node $\boldsymbol{i}_l$, we draw a document-specific beta random variable that acts as a stochastic switch; given a word is at node $\boldsymbol{i}_l$, it determines the probability that the word uses the topic at that node or continues on down the tree. That is, given the path for word $W_{d,n}$ is at node $\boldsymbol{i}_l$, stop with probability

$$U_{d,\boldsymbol{i}_l} \overset{iid}{\sim} \text{Beta}(\gamma_1, \gamma_2), \tag{10}$$

or continue by selecting node $\boldsymbol{i}_{l+1}$ according to $G^{(d)}_{\boldsymbol{i}_l}$. We observe the stick-breaking construction implicit in this construction; for word n in document d, the probability that its topic $\varphi_{d,n} = \theta_{\boldsymbol{i}_l}$ is

$$\Pr(\varphi_{d,n} = \theta_{\boldsymbol{i}_l} | \mathcal{T}_d, \boldsymbol{U}_d) = \left[\prod_{\boldsymbol{i}_m \subset \boldsymbol{i}_l} G^{(d)}_{\boldsymbol{i}_m} \left(\{\theta_{\boldsymbol{i}_{m+1}}\} \right) \right] \left[U_{d,\boldsymbol{i}_l} \prod_{m=1}^{l-1} (1 - U_{d,\boldsymbol{i}_m}) \right]. \tag{11}$$

We use $\boldsymbol{i}_m \subset \boldsymbol{i}_l$ to indicate that the first m values in $\boldsymbol{i}_l$ are equal to $\boldsymbol{i}_m$. The leftmost term in this expression is the probability of path $\boldsymbol{i}_l$, the right term is the probability that the word does not select the first $l-1$ topics, but does select the lth. Since all random variables are independent, a simple product form results that will significantly aid the development of a posterior inference algorithm. The overlapping nature of this stick-breaking construction on the levels of a sequence is evident from the fact that the random variables U are shared for the first l values by all paths along the subtree starting at node $\boldsymbol{i}_l$. A similar tree-structured prior distribution was presented by Adams, et al. [18] in which all groups shared the same distribution on a tree and entire objects (e.g. images or documents) were clustered within a single node. We summarize our model for generating documents with the nHDP in Algorithm 1.

IV. Stochastic variational inference for the Nested HDP

Many text corpora can be viewed as "Big Data"—they are large data sets for which standard inference algorithms can be prohibitively slow. For example, *Wikipedia* currently indexes several million entries, and *The New York Times* has published almost two million articles in the last 20 years. With so much data, fast

inference algorithms are essential. Stochastic variational inference is a development in this direction for hierarchical Bayesian models in which ideas from stochastic optimization are applied to approximate Bayesian inference using mean-field variational Bayes [19][7]. Stochastic inference algorithms have provided significant speed-ups in inference for probabilistic topic models [8][9][20]. In this section, after reviewing the ideas behind stochastic variational inference, we present a stochastic variational inference algorithm for the nHDP topic model.

A. Stochastic variational inference

Stochastic variational inference exploits the difference between *local* variables, or those associated with a single unit of data, and *global* variables, which are shared among an entire data set. In brief, stochastic VB works by splitting a large data set into smaller groups, processing the local variables of one group, updating the global variables, and then moving to another group. This is in contrast to batch inference, which processes all local variables at once before updating the global variables. In the context of probabilistic topic models, the unit of data is a document, and the global variables include the topics (among other variables), while the local variables relate to the distribution on these topics for each document. We next briefly review the relevant ideas from variational inference and its stochastic variant.

1) The batch set-up: Mean-field variational inference is a method for approximate posterior inference in Bayesian models [21]. It approximates the full posterior of a set of model parameters $P(\Phi|W)$ with a factorized distribution $Q(\Phi|\Psi) = \prod_i q_i(\phi_i|\psi_i)$. It does this by searching the space of variational approximations for one that is close to the posterior according to their Kullback-Liebler divergence. Algorithmically, this is done by maximizing the variational objective $\mathcal{L}$ with respect to the variational parameters Ψ of Q, where

$$\mathcal{L}(W, \Psi) = \mathbb{E}_Q[\ln P(W, \Phi)] - \mathbb{E}_Q[\ln Q]. \tag{12}$$

We are interested in conjugate exponential models, where the prior and likelihood of all nodes of the model fall within the conjugate exponential family. In this case, variational inference has a simple optimization procedure [22], which we illustrate with the following example—this generic example gives the general form exploited by the stochastic variational inference algorithm that we apply to the nHDP.

Consider D independent samples from an exponential family distribution $P(W|\eta)$, where η is the natural parameter vector. The likelihood under this model has the standard form

$$P(W_1, \ldots, W_D|\eta) = \left[\prod_{d=1}^{D} h(w_d)\right] \exp\left\{\eta^T \sum_{d=1}^{D} t(w_d) - DA(\eta)\right\}.$$

The sum of vectors $t(w_d)$ forms the sufficient statistics of the likelihood. The conjugate prior on η has a similar form

$$P(\eta|\chi,\nu) = f(\chi,\nu)\exp\left\{\eta^T\chi - \nu A(\eta)\right\}.$$

Conjugacy between these two distributions motivates selecting a q distribution in this same family to approximate the posterior of η,

$$q(\eta|\chi',\nu') = f(\chi',\nu')\exp\left\{\eta^T\chi' - \nu' A(\eta)\right\}.$$

The variational parameters χ' and ν' are free and are modified to maximize the lower bound in Equation (12).[4] Inference proceeds by taking the gradient of $\mathcal{L}$ with respect to the variational parameters of a particular q, in this case the vector $\psi := [\chi'^T, \nu']^T$, and setting to zero to find their updated values. For the conjugate exponential example we are considering, this gradient is

$$\nabla_\psi \mathcal{L}(W,\Psi) = -\begin{bmatrix} \frac{\partial^2 \ln f(\chi',\nu')}{\partial\chi'\partial\chi'^T} & \frac{\partial^2 \ln f(\chi',\nu')}{\partial\chi'\partial\nu'} \\ \frac{\partial^2 \ln f(\chi',\nu')}{\partial\nu'\partial\chi'^T} & \frac{\partial^2 \ln f(\chi',\nu')}{\partial\nu'^2} \end{bmatrix} \begin{bmatrix} \chi + \sum_{d=1}^{D} t(w_d) - \chi' \\ \nu + D - \nu' \end{bmatrix}. \tag{13}$$

Setting this to zero, one can immediately read off the variational parameter updates from the rightmost vector. In this case they are $\chi' = \chi + \sum_{d=1}^{D} t(w_d)$ and $\nu' = \nu + D$, which involve the sufficient statistics for the q distribution calculated from the data.

2) A stochastic extension: Stochastic optimization of the variational lower bound modifies batch inference by forming a noisy gradient of $\mathcal{L}$ at each iteration. The variational parameters for a random subset of the data are optimized first, followed by a step in the direction of the noisy gradient of the global variational parameters. Let $C_s \subset \{1,\ldots,D\}$ index a subset of the data at step s. Also let ϕ_d be the hidden local variables associated with observation w_d and let Φ_W be the global variables shared among all observations. The stochastic variational objective function $\mathcal{L}_s$ is the noisy version of $\mathcal{L}$ formed by selecting a subset of the data,

$$\mathcal{L}_s(W_{C_s},\Psi) = \frac{D}{|C_s|}\sum_{d\in C_s}\mathbb{E}_Q[\ln P(w_d,\phi_d|\Phi_W)] + \mathbb{E}_Q[\ln P(\Phi_W) - \ln Q]. \tag{14}$$

This takes advantage of the conditional independence among the data, and so the log of the joint likelihood can be written as a sum over the D observations. Optimizing $\mathcal{L}_s$ optimizes $\mathcal{L}$ in expectation; since each subset C_s is equally probable, with $p(C_s) = \binom{D}{|C_s|}^{-1}$, and since $d \in C_s$ for $\binom{D-1}{|C_s|-1}$ of the $\binom{D}{|C_s|}$ possible subsets, it follows that $\mathbb{E}_{p(C_s)}[\mathcal{L}_s(W_{C_s},\Psi)] = \mathcal{L}(W,\Psi)$.

[4]A closed form expression for the lower bound is readily derived for this example.

Stochastic variational inference proceeds by optimizing the objective in (14) with respect to ψ_d for $d \in C_s$, followed by an update to Ψ_W that blends the new information with the old. For example, in our conjugate exponential example the update of the global variational parameter $\psi := [\chi'^T, \nu']^T$ at step s is $\psi_s = \psi_{s-1} + \rho_s B \nabla_\psi \mathcal{L}_s(W_{C_s}, \Psi)$, where the matrix B is a positive definite preconditioning matrix and ρ_s is a step size satisfying $\sum_{s=1}^{\infty} \rho_s = \infty$ and $\sum_{s=1}^{\infty} \rho_s^2 < \infty$, which ensures convergence [19].

The gradient $\nabla_\psi \mathcal{L}_s(W_{C_s}, \Psi)$ has a similar form as Equation (13), with the exception that the sum is taken over a subset of the data. Though the matrix in (13) is often very complicated, it is superfluous to batch variational inference for conjugate exponential family models. In the stochastic optimization of Equation (12), however, this matrix cannot be similarly ignored. The key to stochastic variational inference for conjugate exponential models is in selecting B. Since the gradient of $\mathcal{L}_s$ has the same form as Equation (13), B can be set to the inverse of the matrix in (13) to allow for cancellation. An interesting observation is that this matrix is

$$B = -\left(\frac{\partial^2 \ln q(\eta|\psi)}{\partial \psi \partial \psi^T} \right)^{-1}, \tag{15}$$

which is the inverse Fisher information of the variational distribution $q(\eta|\psi)$. Using this B, the step direction is the natural gradient of the lower bound, and therefore not only simplifies the algorithm, but also gives an efficient step direction [23]. The resulting variational update is a ρ_s-weighted combination of the old sufficient statistics for q with the new ones calculated over data indexed by C_s.

B. The inference algorithm

We develop a stochastic variational inference algorithm for approximate posterior inference of the nHDP topic model. As discussed in our general review of stochastic inference, this entails optimizing the local variational parameters for a subset of documents, followed by a step along the natural gradient of the global variational parameters. We distinguish between local and global variables for the nHDP in Table II. In Table II we also give the variational q distributions selected for each variable. In almost all cases, we select this distribution to be in the same family as the prior. We point out two additional latent indicator variables that we have added for inference: $c_{d,n}$, which indicates the topic from which $W_{d,n}$ is drawn, and $z_{\boldsymbol{i},j}^{(d)}$, which points to the atom in $G_{\boldsymbol{i}}$ for the jth break in $G_{\boldsymbol{i}}^{(d)}$ using the construction given in (9).

In addition to local and global variational parameter updates, we introduce a third aspect to our inference algorithm. Before optimizing any variational parameters, we select a subtree from $\mathcal{T}$ for each document using a greedy algorithm. This greedy algorithm is performed with respect to the variational objective

TABLE I

A LIST OF THE LOCAL AND GLOBAL VARIABLES AND THEIR RESPECTIVE q DISTRIBUTIONS FOR THE NHDP TOPIC MODEL.

Global variables:	$\theta_{\boldsymbol{i}}$: topic probability vector for node $\boldsymbol{i}$	$q(\theta_{\boldsymbol{i}}) = \text{Dirichlet}(\theta_{\boldsymbol{i}} \lvert \lambda_{\boldsymbol{i},1}, \ldots, \lambda_{\boldsymbol{i},\mathcal{V}})$
	$V_{\boldsymbol{i},j}$: stick proportion for the top-level DP for node $\boldsymbol{i}$	$q(V_{\boldsymbol{i},j}) = \text{Beta}(V_{\boldsymbol{i},j} \lvert \tau_{\boldsymbol{i},j}^{(1)}, \tau_{\boldsymbol{i},j}^{(2)})$
Local variables:	$V_{\boldsymbol{i},j}^{(d)}$: stick proportion for second-level DP for node $\boldsymbol{i}$	$q(V_{\boldsymbol{i},j}) = \text{Beta}(V_{\boldsymbol{i},j}^{(d)} \lvert u_{\boldsymbol{i},j}^{(d)}, v_{\boldsymbol{i},j}^{(d)})$
	$z_{\boldsymbol{i},j}^{(d)}$: index pointer to atom in $G_{\boldsymbol{i}}$ for jth break in $G_{\boldsymbol{i}}^{(d)}$	$q(z_{\boldsymbol{i},j}^{(d)}) = \delta_{z_{\cdot,\cdot}^{\cdot}}(k),\ k = 1, 2, \ldots$
	$U_{d,\boldsymbol{i}}$: beta distributed switch probability for node $\boldsymbol{i}$	$q(U_{d,\boldsymbol{i}}) = \text{Beta}(U_{d,\boldsymbol{i}} \lvert a_{d,\boldsymbol{i}}, b_{d,\boldsymbol{i}})$
	$c_{d,n}$: topic indicator for word n in document d	$q(c_{d,n}) = \text{Discrete}(c_{d,n} \lvert \nu_{d,n})$

function, and so we are still performing variational inference. This limits the number of paths for which variational parameters must be learned for a given document, which further speeds up inference. We discuss this greedy algorithm below, followed by the variational parameter updates for the local and global q distributions.

1) Greedy subtree selection: As mentioned, we perform a greedy algorithm with respect to the variational objective function to determine a subtree from $\mathcal{T}$ for each document. We first describe the algorithm followed by a mathematical representation. Starting from the root node, we sequentially add nodes from $\mathcal{T}$ from those currently "activated." An activated node is one whose parent is contained within the subtree but which is not itself in the subtree. We hold the q distributions for the document-specific beta distributions fixed to their priors and set the variational distribution for each word's topic indicator to zero on all unactivated nodes. We then ask: Which of the activated nodes not currently in the subtree will lead to the greatest increase in the variational objective? This only involves optimizing the variational parameter for each word over the current subtree plus the candidate node, which does not require iterating. We continue adding the maximizing node until the marginal increase in the objective falls below a threshold. We formalize this process below.

a) Coordinate update for $q(z_{\boldsymbol{i},j}^{(d)})$: As defined in Table II, $z_{\boldsymbol{i},j}^{(d)}$ is the variable that indicates the index of the atom from the top-level DP $G_{\boldsymbol{i}}$ pointed to by the jth stick-breaking weight in $G_{\boldsymbol{i}}^{(d)}$. We select a delta q distribution for this variable, meaning we make a hard assignment for this value. Starting with an empty tree, all atoms in $G_{\boldsymbol{i}_0}$ constitute the activated set. Adding the first node is equivalent to determining the value for $z_{\boldsymbol{i}_0,1}^{(d)}$; in general, creating a subtree for $\mathcal{T}_d$, denoted $\mathcal{T}_d'$, is equivalent to determining which $z_{\boldsymbol{i},j}^{(d)}$ to include in $\mathcal{T}_d'$ and the atoms to which they point.

For a subtree of size t corresponding to document d, let the set $\mathcal{I}_{d,t}$ contain the index values of the included nodes, let $\mathcal{S}_{d,t} = \{\boldsymbol{i} : pa(\boldsymbol{i}) \in \mathcal{I}_{d,t}, \boldsymbol{i} \notin \mathcal{I}_{d,t}\}$. Also, let $\mathcal{C}_{d,t,\boldsymbol{i}'}$ denote the conditions that

$\nu_{d,n}(\boldsymbol{i}) = 0$ for all $\boldsymbol{i} \notin \mathcal{I}_{d,t} \cup \boldsymbol{i}'$ and that $q(\cdot)$ is set fixed to the prior for all other document specific distributions. Then provided the marginal increase in the variational objective is above a preset threshold, we increment the subtree by $\mathcal{I}_{d,t+1} \leftarrow \mathcal{I}_{d,t} \cup \boldsymbol{i}^*$, where

$$\boldsymbol{i}^* = \arg\max_{\boldsymbol{i}' \in \mathcal{S}_{d,t}} \sum_{n=1}^{N_d} \max_{\nu_{d,n}: \mathcal{C}_{d,t,\boldsymbol{i}'}} \mathbb{E}_q[\ln p(w_{d,n}|c_{d,n},\theta)] + \mathbb{E}_q[\ln p(c_{d,n}, \boldsymbol{z}^{(d)}|V, V_d, U_d)] - \mathbb{E}_q[\ln q(c_{d,n})]. \quad (16)$$

The optimal values for $\nu_{d,n}$ are given below in Equation (17). We note two aspects of this greedy algorithm. First, though the stick-breaking construction of the second-level DP given in (9) allows for atoms to repeat, in this algorithm each added atom is new, since there is no advantage in duplicating atoms. Therefore, the algorithm approximates each $G_{\boldsymbol{i}}^{(d)}$ by selecting and reordering a subset of atoms from $G_{\boldsymbol{i}}$ for its stick-breaking construction. (The subtree $\mathcal{T}_d'$ may contain no atoms or one atom from a $G_{\boldsymbol{i}}$.) The second aspect we point out is the changing prior on the same node in $\mathcal{T}$. If the atom $\theta_{(\boldsymbol{i},m)}$ is a candidate for addition, then it remains a candidate until it is either selected by a $z_{\boldsymbol{i},j}^{(d)}$, or the algorithm terminates. The prior on selecting this atom changes, however, depending on whether it is a candidate for $z_{\boldsymbol{i},j}^{(d)}$ or $z_{\boldsymbol{i},j'}^{(d)}$. Therefore, incorporating a sibling of $\theta_{(\boldsymbol{i},m)}$ impacts the prior on incorporating $\theta_{(\boldsymbol{i},m)}$. This penalty corresponds to the prior on word indicators, and is in addition to the penalty of the atom itself from the top-level DP.

2) Coordinate updates for document variables: Given the subtree $\mathcal{T}_d'$ selected for document d, we optimize the variational parameters for the q distributions on $c_{d,n}$, $V_{\boldsymbol{i},j}^{(d)}$ and $U_{d,\boldsymbol{i}}$ over that subtree.

a) Coordinate update for $q(c_{d,n})$: The variational distribution on the path for word $W_{d,n}$ is

$$\nu_{d,n}(\boldsymbol{i}) \propto \exp\left\{\mathbb{E}_q[\ln \theta_{\boldsymbol{i},W_{d,n}}] + \mathbb{E}_q[\ln \pi_{d,\boldsymbol{i}}]\right\}, \quad (17)$$

where the prior term $\pi_{d,\boldsymbol{i}}$ is the tree-structured prior of the nHDP,

$$\pi_{d,\boldsymbol{i}} = \left[\prod_{(\boldsymbol{i}',i)\subseteq \boldsymbol{i}} \prod_j \left(V_{\boldsymbol{i}',j}^{(d)} \prod_{m<j}(1 - V_{\boldsymbol{i}',m}^{(d)})\right)^{\mathbb{I}(z_{\boldsymbol{i}',j}^{(d)}=i)}\right] \left[U_{d,\boldsymbol{i}} \prod_{\boldsymbol{i}' \subset \boldsymbol{i}} (1 - U_{d,\boldsymbol{i}'})\right]. \quad (18)$$

The expectation $\mathbb{E}_q[\ln \theta_{\boldsymbol{i},w}] = \psi(\lambda_{\boldsymbol{i},w}) - \psi(\sum_w \lambda_{\boldsymbol{i},w})$, where $\psi(\cdot)$ is the digamma function. Similarly, for a random variable $Y \sim \text{Beta}(a,b)$, $\mathbb{E}[\ln Y] = \psi(a) - \psi(a+b)$ and $\mathbb{E}[\ln(1-Y)] = \psi(b) - \psi(a+b)$. The corresponding values of a and b for U and V are given in their respective updates below.

We note that, given the subtree of $\mathcal{T}_d$ the distribution on paths has a familiar feel as LDA, but where LDA uses a flat Dirichlet prior on π_d, the nHDP uses a prior that is the product of several beta random variables having a tree-structured form. Though the form is more complicated, the independence results in simple closed-form updates for these beta variables that only depend on $\nu_{d,n}$.

b) Coordinate update for $q(V_{\boldsymbol{i},j}^{(d)})$: The variational parameter updates for the second-level stick-breaking proportions are

$$u_{\boldsymbol{i},j}^{(d)} = 1 + \sum_{\boldsymbol{i}':(\boldsymbol{i},j)\subseteq\boldsymbol{i}'} \textstyle\sum_{n=1}^{N_d} \nu_{d,n}(\boldsymbol{i}'), \tag{19}$$

$$v_{\boldsymbol{i},j}^{(d)} = \beta + \sum_{\boldsymbol{i}':\boldsymbol{i}\subset\boldsymbol{i}'} \mathbb{I}(\cup_{m>j}\{z_{\boldsymbol{i},m}^{(d)} = \boldsymbol{i}'(l+1)\}) \textstyle\sum_{n=1}^{N_d} \nu_{d,n}(\boldsymbol{i}'). \tag{20}$$

The statistic for the first parameter is the expected number of words in document d that pass through or stop at node $(\boldsymbol{i}, j)$. The statistic for the second parameter is the expected number of words from document d whose paths pass through the same parent $\boldsymbol{i}$, but then transition to a node with index greater than j according to the indicators $z_{\boldsymbol{i},m}^{(d)}$ from the second-level stick-breaking construction of $G_{\boldsymbol{i}}^{(d)}$.

c) Coordinate update for $q(U_{d,\boldsymbol{i}})$: The variational parameter updates for the switching probabilities are similar to those of the second-level stick-breaking process, but collect the statistics from $\nu_{d,n}$ in a slightly different way,

$$a_{d,\boldsymbol{i}} = \gamma_1 + \textstyle\sum_{n=1}^{N_d} \nu_{d,n}(\boldsymbol{i}), \tag{21}$$

$$b_{d,\boldsymbol{i}} = \gamma_2 + \sum_{\boldsymbol{i}':\boldsymbol{i}\subset\boldsymbol{i}'} \textstyle\sum_{n=1}^{N_d} \nu_{d,n}(\boldsymbol{i}'). \tag{22}$$

The statistic for the first parameter is the expected number of words that use the topic at node $\boldsymbol{i}$. The statistic for the second parameter is the expected number of words that pass through node $\boldsymbol{i}$ but do not terminate there.

3) Stochastic updates for corpus variables: After selecting the subtrees and updating the local document-specific variational parameters for each document d in sub-batch s, we take a step in the direction of the natural gradient of the parameters of the q distributions on the global variables. These include the topics $\theta_{\boldsymbol{i}}$ and the top-level stick-breaking proportions $V_{\boldsymbol{i}_l,j}$.

a) Stochastic update for $q(\theta_{\boldsymbol{i}})$: For the stochastic update of the Dirichlet q distributions on each topic $\theta_{\boldsymbol{i}}$, first form the vector $\lambda_{\boldsymbol{i}}'$ of sufficient statistics using the data in sub-batch s,

$$\lambda_{\boldsymbol{i},w}' = \frac{D}{|C_s|} \sum_{d\in C_s} \textstyle\sum_{n=1}^{N_d} \nu_{d,n}(\boldsymbol{i})\mathbb{I}\{W_{d,n} = w\}, \quad w = 1,\ldots,\mathcal{V}.$$

This vector contains the expected count of the number of words with index w that originate from topic $\theta_{\boldsymbol{i}}$ over documents indexed by C_s. According to the stochastic inference theory in Section IV-A2, this number is scaled up to a corpus of size D. The update to the variational parameters for the associated q distribution is

$$\lambda_{\boldsymbol{i},w}^{s+1} = \lambda_0 + (1-\rho_s)\lambda_{\boldsymbol{i},w}^{s} + \rho_s\lambda_{\boldsymbol{i},w}'. \tag{23}$$

We see a blending of the old with the new in this update. Since $\rho_s \to 0$ as s increases, the algorithm uses less and less information from new sub-groups of documents, which reflects the increasing confidence in this parameter value as more data is seen.

b) Stochastic update for $q(V_{i_l,j})$: Similarly to $\theta_{\boldsymbol{i}}$, we first collect the sufficient statistics for the q distribution on $\boldsymbol{V}_{\boldsymbol{i}_l,j}$ from the documents in sub-batch s,

$$\tau'_{\boldsymbol{i}_l,j} = \frac{D}{|C_s|} \sum_{d \in C_s} \mathbb{I}\{\boldsymbol{i}_l \in \mathcal{I}_d\}, \quad \tau''_{\boldsymbol{i}_l,j} = \frac{D}{|C_s|} \sum_{d \in C_s} \sum_{j > i_l} \mathbb{I}\{(pa(\boldsymbol{i}_l), j) \in \mathcal{I}_d\}.$$

The first value scales up the number of documents in sub-batch s that include atom $\theta_{(\boldsymbol{i},j)}$ in their subtree; the second value scales up the number of times an atom of higher index value in the same DP is used by a document in sub-batch s. The update to the global variational parameters are

$$\tau^{(1)}_{\boldsymbol{i}_l,j}(s+1) = 1 + (1-\rho_s)\tau^{(1)}_{\boldsymbol{i}_l,j}(s) + \rho_s \tau'_{\boldsymbol{i}_l,j}, \tag{24}$$

$$\tau^{(2)}_{\boldsymbol{i}_l,j}(s+1) = \alpha + (1-\rho_s)\tau^{(2)}_{\boldsymbol{i}_l,j}(s) + \rho_s \tau''_{\boldsymbol{i}_l,j}. \tag{25}$$

Again, we see a blending of old information with new.

V. Experiments

We present an empirical evaluation of the nested HDP topic model in the stochastic and the batch inference settings. We first present batch results on three smaller data sets to verify that our multi-path approach gives an improvement over the single-path nested CRP. We then move to the stochastic inference setting, where we perform experiments on 1.8 million documents from *The New York Times* and 3.3 million documents from *Wikipedia*. We compare with other recent stochastic inference algorithms for topic models: stochastic LDA [8] and the stochastic HDP [9]. Before presenting our results, we discuss our method for initializing the topic q distributions of the tree.

A. Initialization

As with most Bayesian models, inference for hierarchical topic models can benefit greatly from a good initialization. Our goal is to find a method for quickly centering the posterior mean of each topic so that they contain some information about their hierarchical relationships. We briefly discuss our approach for initializing the global variational topic parameters $\lambda_{\boldsymbol{i}}$ of the nHDP.

Using a small set of documents from the training set, we form the empirical distribution for each document on the vocabulary. We then perform k-means clustering of these probability vectors using the L_1 distance measure. At the top level, we partition the data into n_1 groups, corresponding to n_1 children

of the root node. We then subtract the mean of a group (a probability vector) from all data within that group, set any negative values to zero and renormalize. We loosely think of this as the "probability of what remains"—a distribution on words not captured by the parent distributions. Within each group we again perform k-means clustering, obtaining n_2 probability vectors for each of the n_1 groups, and again subtracting, setting negative values to zero and renormalizing the remainder of each probability vector for a document.

Through this hierarchical k-means clustering, we obtain n_1 probability vectors at the top level, n_2 probability vectors beneath each top-level vector for the second level, n_3 probability vectors beneath each of these second-level vectors, etc. The n_i vectors obtained from any sub-group of data are refinements of an already coherent sub-group of data, since that sub-group is itself a cluster from a larger group. Therefore, the resulting tree will have some thematic coherence. The clusters from this algorithm parallel the nodes within the nHDP tree. For a mean probability vector $\hat{\lambda}_{\boldsymbol{i}}$ obtained from this algorithm, we set the corresponding variational parameter for the topic Dirichlet q distribution to $\lambda_{\boldsymbol{i}} = N(\rho\hat{\lambda}_{\boldsymbol{i}} + (1-\rho)\mathbf{1}/\mathcal{V})$ for $\rho \in [0, 1]$ and N a scaling factor. This initializes the mean of $\theta_{\boldsymbol{i}}$ to be slightly peaked around $\hat{\lambda}_{\boldsymbol{i}}$, while the uniform vector and ρ determine the variance. In our algorithms we set $\rho = 0.5$ and N equal to the number of documents.

B. A batch comparison

Before comparing our stochastic inference algorithm for the nHDP with similar algorithms for LDA and the HDP, we compare a batch version with the nCRP on three smaller data sets. This will verify the advantage of giving each document access to the entire tree versus forcing each document to follow one path. We compare the variational nHDP topic model with both the variational nCRP [2] and the Gibbs sampling nCRP [1]. We consider three corpora for our experiments: (i) The *Journal of the ACM*, a collection of 536 abstracts from the years 1987–2004 with vocabulary size 1,539; (ii) The *Psychological Review*, a collection of 1,272 abstracts from the years 1967–2003 with vocabulary size 1,971; and (iii) The *Proceedings of the National Academy of Science*, a collection of 5,000 abstracts from the years 1991–2001 with a vocabulary size of 7,762. The average number of words per document for the three corpora are 45, 108 and 179, respectively.

Variational inference for Dirichlet priors uses a truncation of the variational distribution, which limits the number of topics that are learned [24][25]. This truncation is set to a number larger than the anticipated number of topics necessary for modeling the data set, but can adapt if more are needed [26]. We use a truncated tree of $(10, 7, 5)$ for modeling these corpora, where 10 children of the root node each have 7

TABLE II

COMPARISON OF THE NHDP WITH THE NCRP ON THREE SMALLER PROBLEMS.

Method\Data set	JACM	Psych. Review	PNAS
Variational nHDP	-5.405 $\pm$ 0.012	-5.674 $\pm$ 0.019	-6.304 $\pm$ 0.003
Variational nCRP	-5.433 $\pm$ 0.010	-5.843 $\pm$ 0.015	-6.574 $\pm$ 0.005
Gibbs nCRP	-5.392 $\pm$ 0.005	-5.783 $\pm$ 0.015	-6.496 $\pm$ 0.007

children, which themselves each have 5 children for a total of 420 nodes. Following previous work on the nCRP, we truncate the tree to three levels. Also, because these three data sets contain stop words, we follow [2] and [1] by including a root node shared by all documents for this batch problem. Following [2], we perform five-fold cross validation to evaluate performance on each corpora.

We present our results in Table II. We see that for all data sets, the variational nHDP outperforms the variational nCRP. For the two larger data sets, the variational nHDP also outperforms Gibbs sampling for the nCRP. Given the relative sizes of these corpora, we see that the benefit of learning a per-document distribution on the full tree rather than a path appears to increase as the corpus size and document size increase. Since we are interested in the "Big Data" regime, this strongly hints at an advantage of our nHDP approach over the nCRP.

C. Stochastic inference for The New York Times and Wikipedia

We next present an evaluation of our stochastic variational inference algorithm on *The New York Times* and *Wikipedia*. These are both very large data sets, with *The New York Times* containing roughly 1.8 million articles and *Wikipedia* roughly 3.3 million web pages. The average document size is somewhat larger than those considered in our batch experiments as well, with an article from *The New York Times* containing 254 words on average taken from a vocabulary size of 8,000, and *Wikipedia* 164 words on average taken from a vocabulary size of 7,702.

1) Setup: We use the algorithm discussed in Section V-A to initialize a three-level tree with $(20, 10, 5)$ child nodes per level, giving a total of 1,220 initial topics. For the Dirichlet processes, we set all top-level DP concentration parameters to $\alpha = 5$ and the second-level DP concentration parameters to $\beta = 1$. For the switching probabilities U, we set the beta distribution hyperparameters to $\gamma_1 = 2/3$ and $\gamma_2 = 4/3$, which takes the weight of a uniform prior and skews it toward smaller values, slightly encouraging a word to continue down the tree. For our greedy subtree selection algorithm, we stop adding nodes to the subtree when the marginal improvement to the lower bound falls below 10^{-2}. When optimizing the local variational parameters of a document given its subtree, we continue iterating until the absolute change

in the empirical distribution of words on the tree falls below 10^{-1}.

We hold out a data set for each corpus for testing; we hold out 14,268 documents for testing *The New York Times* and 8,704 documents for testing *Wikipedia*. We quantitatively assess the quality of the tree at any given point in the algorithm as follows: Holding the top-level variational parameters fixed, for each test document we randomly partition the words into a 75/25 percent split. We then learn document-specific variational parameters for the 75% portion. Following [27][2], we use the mean of each q distribution to form a predictive distribution for the remaining words of that document. With this distribution, we calculate the average per-word log likelihood of the 25% portion to assess performance. For comparison, we evaluate stochastic inference algorithms for LDA and the HDP in the same manner. In all algorithms, we use a sub-batch size of $|C_s| = 5000$ at step s and set the learning rate to $\rho_s = (1 + s)^{-0.75}$. We note that Hoffman, et al. [7] provide a detailed evaluation of these settings, and while performance depends on their values, relative performance remains consistent; these settings are in the good performance range according to their evaluation and our qualitative results support this conclusion on these data sets.

2) The New York Times: We first present our results for *The New York Times*. In Figure 2 we show the log likelihood on the test set as a function of number of documents seen by the model. We see an improvement in all algorithms as the amount of data seen increases. We also note an improvement in the performance of the nHDP compared with LDA and the HDP. In Figure 3 we show document-level statistics from the test set at the final step of the algorithm. These include the sizes of the subtrees, a breakdown by level of these subtrees, and word allocations by level. We note that while the tree has three levels, roughly eight topics are being used (in varying degrees) per document. This is in contrast to the three topics that would be available to any document with the nCRP. Thus there is a clear advantage in allowing each document to have access to the entire tree.

In Figure 4 we show example topics from the model and their relative structure. We show four topics from the top level of the tree (shaded), and connect topics according to parent/child relationship. The model learns a meaningful hierarchical structure; for example, the sports subtree branches into the various sports, which themselves appear to branch by teams. In the foreign affairs subtree, children tend to group by major subregion and then branch out into subregion or issue. In Figure 5a we give a sense of the size of the tree as a function of documents seen. Since all topics aren't used equally, we show the number of nodes containing 90%, 99% and 99.9% of all paths within the subtrees.

3) Wikipedia: We find similar results for *Wikipedia* as for *The New York Times*. In Figures 6 and 7 we show results corresponding to Figures 2 and 3 for *The New York Times*. We again see an improvement in performance for the nHDP over LDA and the HDP, as well as the increased usage of the tree with

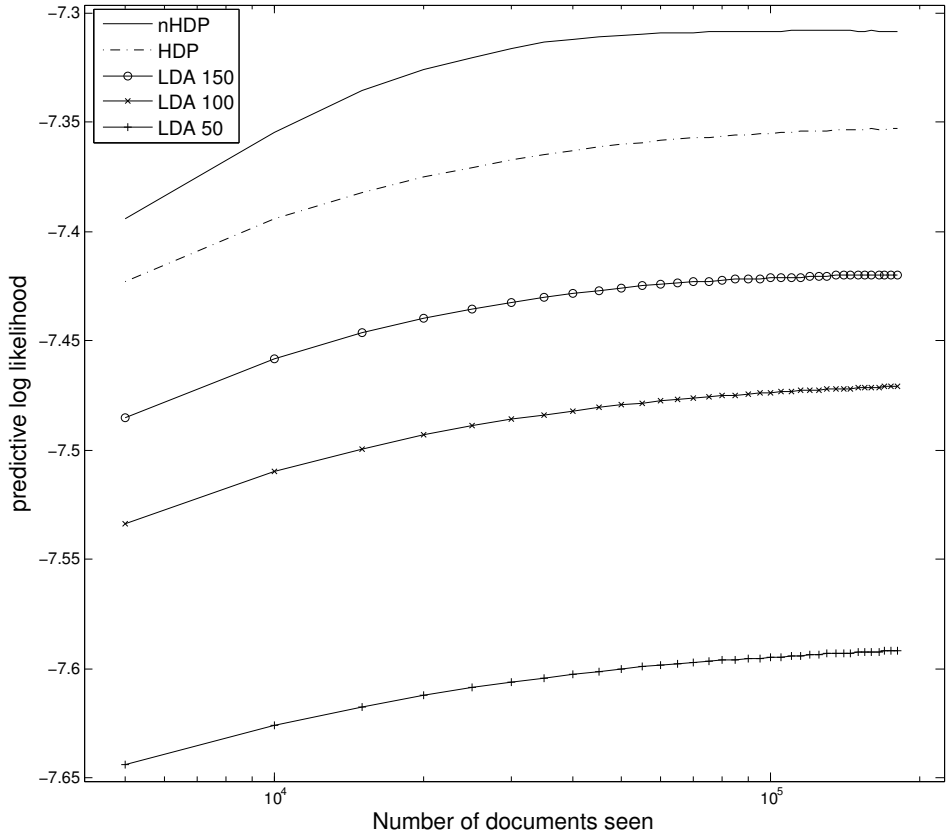

Fig. 2. The New York Times: Average per-word log likelihood on a held-out test set as a function of training documents seen.

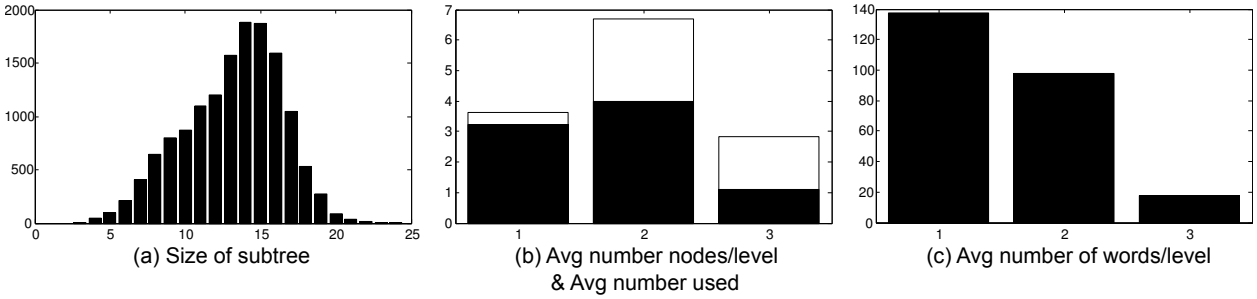

Fig. 3. The New York Times: Per-document statistics from the test set using the tree at the final step of the algorithm. (a) A histogram of the size of the subtree selected for a document. (b) The average number of nodes by level within the subtree (white), and the average number by level that have at least one expected observation (black). (c) The average number of words allocated to each level of the tree per document.

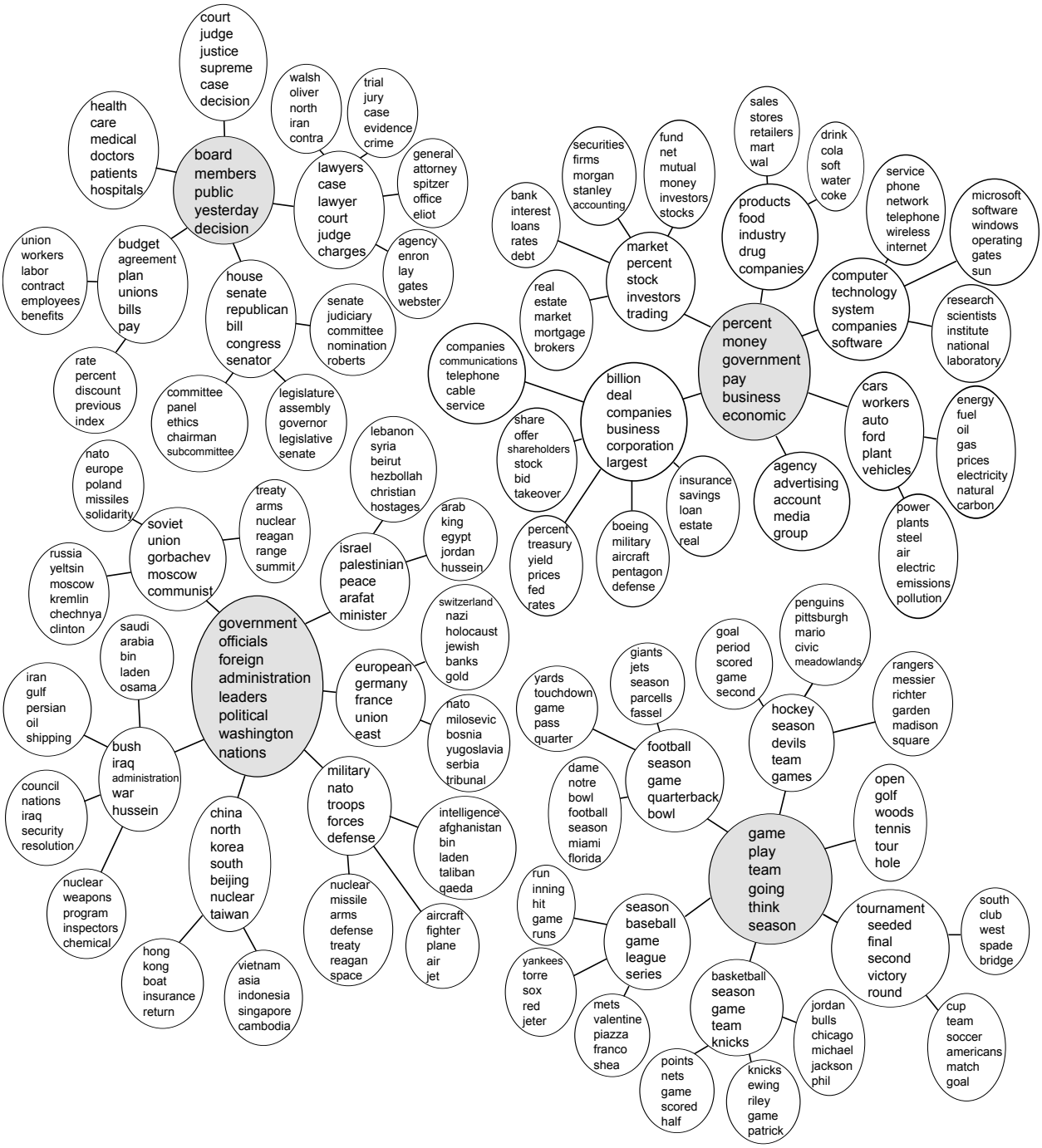

Fig. 4. Tree-structured topics from The New York Times. The shaded node is the top-level node and lines indicate dependencies within the tree. In general, topics are learning in increasing levels of specificity. For clarity, we have removed grammatical variations of the same word, such as “scientist” and “scientists.”

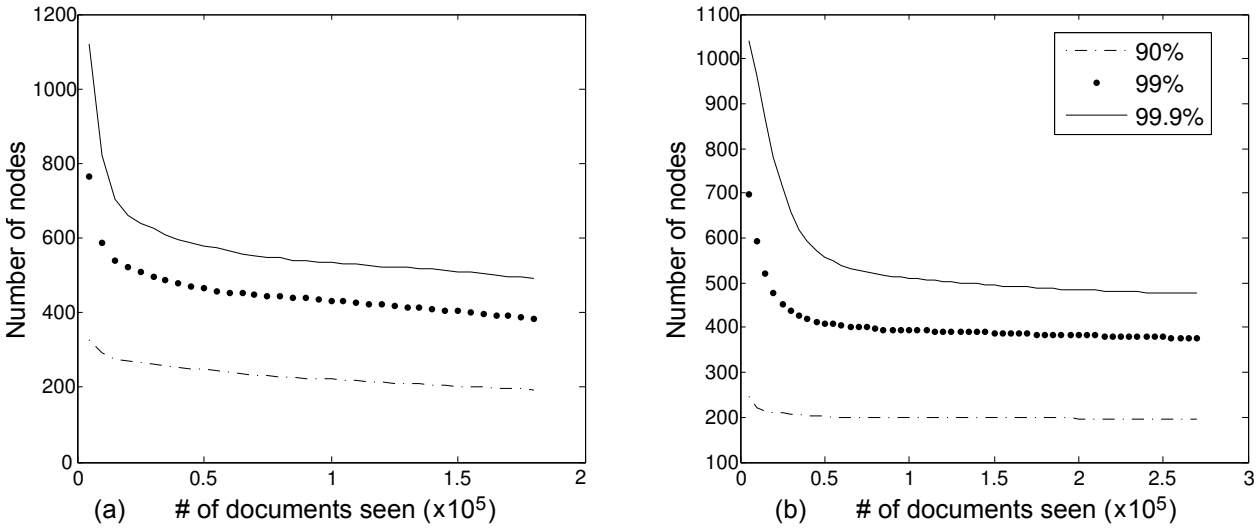

Fig. 5. Tree size: The smallest number of nodes containing 90%, 99% and 99.9% of all paths as a function of documents seen for (a) The New York Times, and (b) Wikipedia.

the nHDP than would be available in the nCRP. In Figure 8, we see example subtrees used by three documents. We note that the topics contain many more function words than for *The New York Times*, but an underlying hierarchical structure is uncovered that would be unlikely to arise along one path, as the nCRP would require. In Figure 5b we again show the size of the tree as a function of documents seen by showing the number of nodes containing 90%, 99% and 99.9% of all paths from the subtrees. As with *The New York Times*, the model simplifies significantly from the original 1,220 nodes initialized.

VI. Conclusion

We have presented the nested hierarchical Dirichlet process (nHDP), an extension of the nested Chinese restaurant process (nCRP) that allows each observation to follow its own path to a topic in the tree. Starting with a stick-breaking construction for the nCRP, the new model samples document-specific path distributions for a shared tree using a hierarchy of Dirichlet processes. By giving a document access to the entire tree, we are able to borrow thematic content from various parts of the tree in constructing a document. In our experiments we showed that this led to a general improvement over the nCRP for hierarchical topic modeling. In addition, we have developed a stochastic variational inference algorithm that is scalable to very large data sets. We compared the stochastic nHDP topic model with stochastic LDA and HDP on large collections from *The New York Times* and *Wikipedia*, where we showed an improvement in predictive ability with our tree-structured prior. Qualitative results on these corpora indicate that the

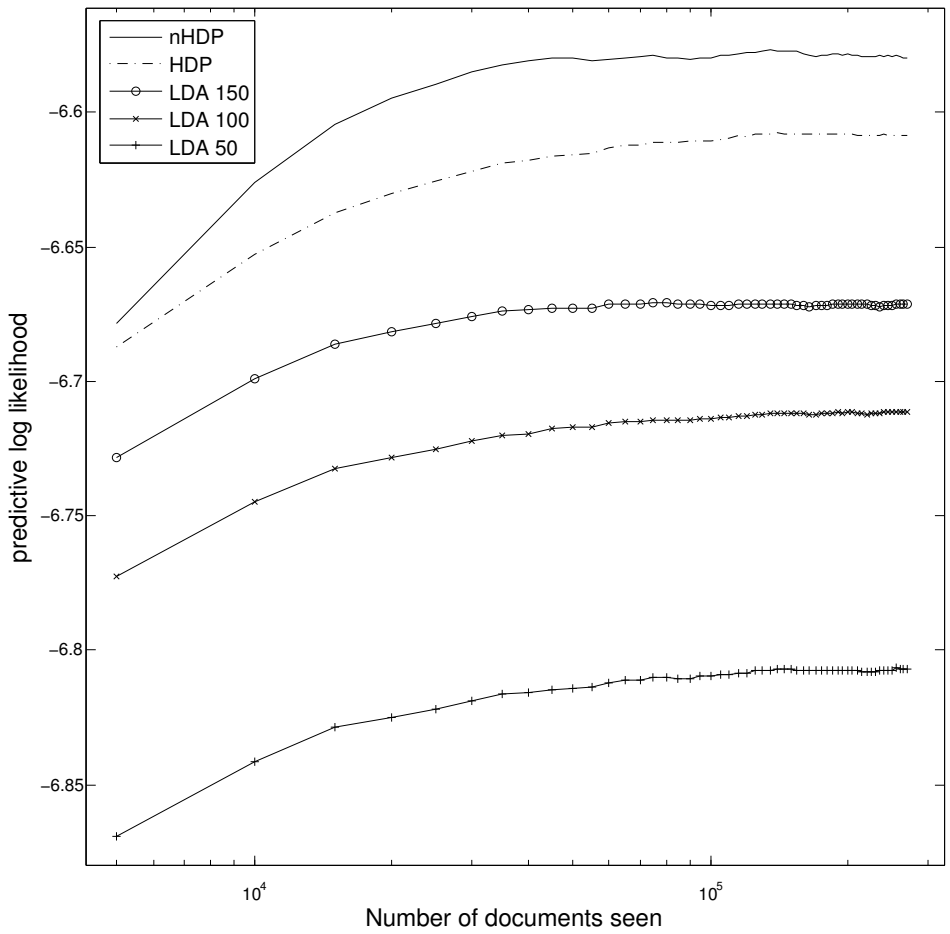

Fig. 6. Wikipedia: Average per-word log likelihood on a held-out test set as a function of training documents seen.

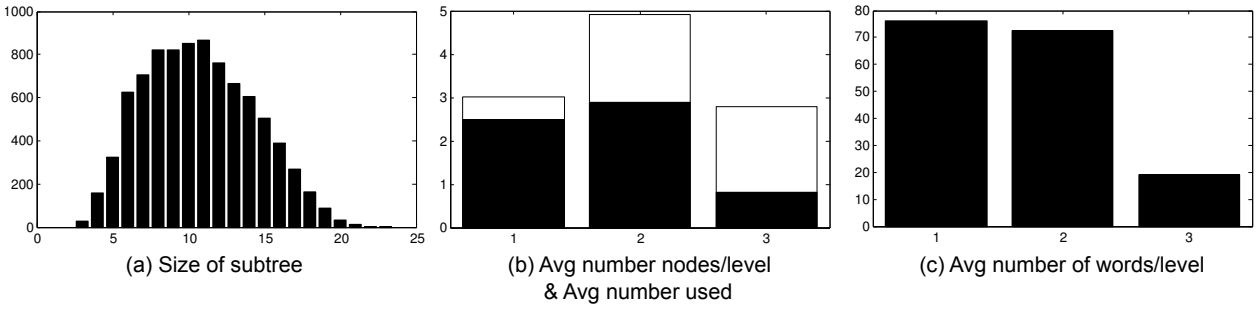

Fig. 7. Wikipedia: Per-document statistics from the test set using the tree at the final step of the algorithm. (a) A histogram of the size of the subtree selected for a document. (b) The average number of nodes by level within the subtree (white), and the average number by level that have at least one expected observation (black). (c) The average number of words allocated to each level of the tree per document.

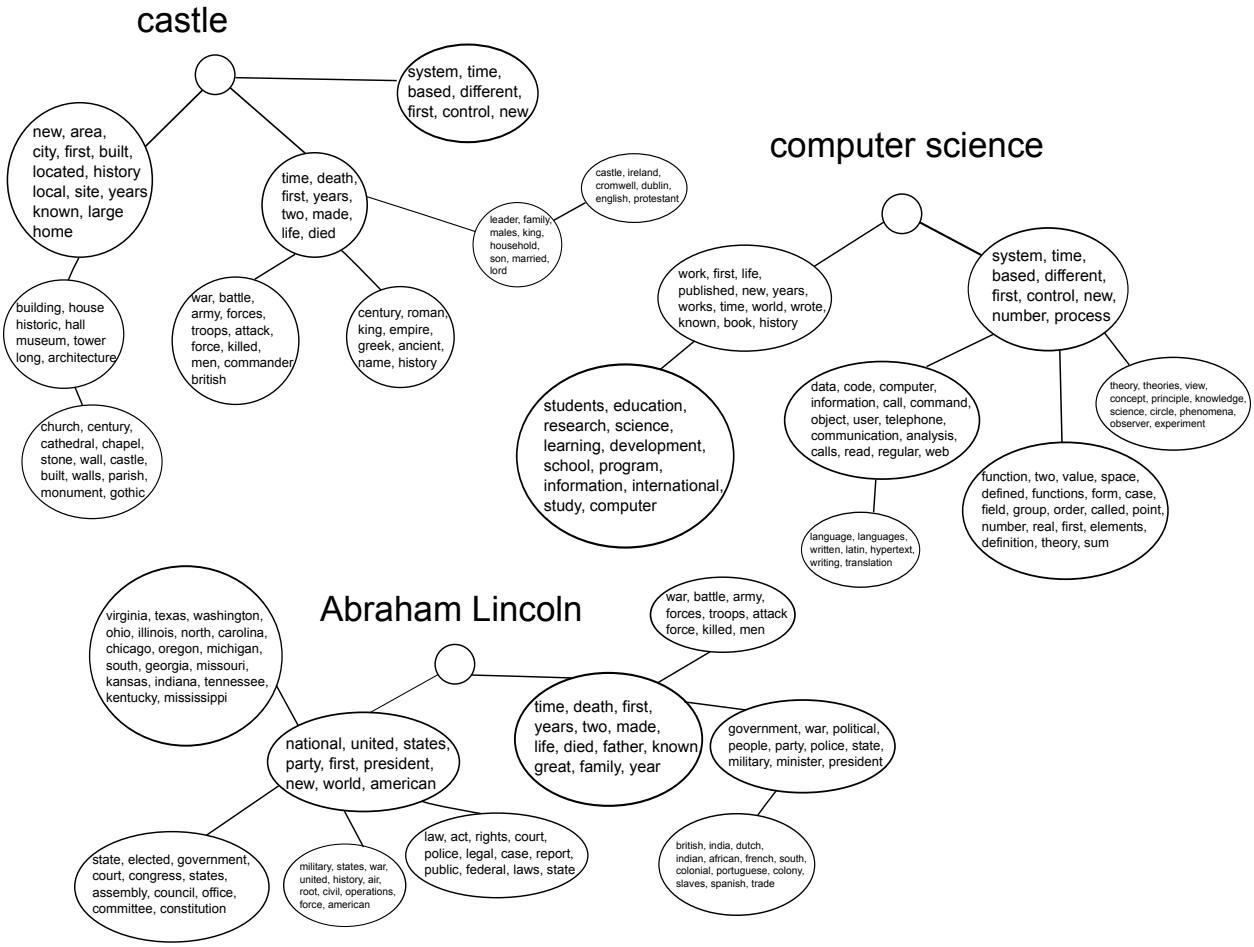

Fig. 8. Examples of subtrees for three articles from *Wikipedia*. The three sizes of font indicate differentiate the more probable topics from the less probable.

nHDP can learn meaningful topic hierarchies, and that documents benefit by taking advantage of the entire tree.

REFERENCES

[1] D. Blei, T. Griffiths, and M. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 7:1–30, 2010.

[2] C. Wang and D. Blei, "Variational inference for the nested Chinese restaurant process," in *Advances in Neural Information Processing Systems*, 2009.

[3] J. H. Kim, D. Kim, S. Kim, and A. Oh, "Modeling topic hierarchies with the recursive Chinese restaurant process," in *International Conference on Information and Knowledge Management (CIKM)*, 2012.

[4] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[5] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[6] M. Jordan, "Message from the President: The era of Big Data," *ISBA Bulletin*, vol. 18, no. 2, pp. 1–3, 2011.

[7] M. Hoffman, D. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *arXiv:1206.7051*, 2012.

[8] M. Hoffman, D. Blei, and F. Bach, "Online learning for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems*, 2010.

[9] C. Wang, J. Paisley, and D. Blei, "Online learning for the hierarchical Dirichlet process," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 15, 2011, pp. 752–760.

[10] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[11] D. Blackwell and J. MacQueen, "Ferguson distributions via Pólya urn schemes," *Annals of Statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[12] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[13] D. Aldous, *Exchangeability and Related Topics*, ser. Ecole d'Eté Probabilités de Saint-Flour XIII-1983 pages 1-198. Springer, 1985.

[14] A. Rodriguez, D. Dunson, and A. Gelfand, "The nested Dirichlet process," *Journal of the American Statistical Association*, vol. 103, pp. 1131–1154, 2008.

[15] L. Ren, L. Carin, and D. Dunson, "The dynamic hierarchical Dirichlet process," in *International Conference on Machine Learning*, 2008.

[16] E. Airoldi, D. Blei, S. Fienberg, and E. Xing, "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, vol. 9, pp. 1981–2014, 2008.

[17] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "A Sticky HDP-HMM with Application to Speaker Diarization," *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[18] R. Adams, Z. Ghahramani, and M. Jordan, "Tree-structured stick breaking for hierarchical data," in *Advances in Neural Information Processing Systems*, 2010.

[19] M. Sato, "Online model selection based on the variational Bayes," *Neural Computation*, vol. 13, no. 7, pp. 1649–1681, 2001.

[20] J. Paisley, C. Wang, and D. Blei, "The discrete infinite logistic normal distribution," *Bayesian Analysis*, vol. 7, no. 2, pp. 235–272, 2012.

[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999.

[22] J. Winn and C. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.

[23] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[24] D. Blei and M. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2005.

[25] K. Kurihara, M. Welling, and N. Vlassis, "Accelerated variational DP mixture models," in *Advances in Neural Information Processing Systems 19*, 2006, pp. 761–768.

[26] C. Wang and D. Blei, "Truncation-free online variational inference for Bayesian nonparametric models," in *Advances in Neural Information Processing Systems*, 2012.

[27] Y. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for HDP," in *Advances in Neural Information Processing Systems*, 2008.